Supplementary Appendix: Testing for Negative Spillovers: Is Human Rights Promotion Really Part of the ‘Problem’

1 A sensitivity analysis for counteraction hypotheses

In the paper, we outline a framework for connecting theories about counteraction in human rights, mediation effects, and instrumental variables estimators. This section further develops the theory connecting the IV estimator to mediation effects, explains the assumptions behind using the IV sensitivity analysis for assessing mediation effects, and distinguishes the proposed approach from existing methods for mediation analysis. In particular, we build on results from Sobel 2008 to show how the IV estimator identifies the path effect of the mediator on the outcome under the exclusion restriction. We allow researchers to relax the exclusion restriction by specifying beliefs about how large the violation of the might be in terms of an intuitive quantity: the expected direct effect of the treatment on the outcome.

1.1 Notation

As in Section III, we define an observed treatment A_i, mediator/targeted outcome M_i, and final/"spillover" outcome Y_i. The potential outcome $M_i(a)$ denotes the mediator that would be observed if unit i were assigned treatment level a. Likewise, the potential outcome $Y_i(a, m)$ is the value of Y_i observed if i were assigned treatment a and mediator value m. We can also write a composite potential outcome $Y_i(a) = Y_i(a, M_i(a))$ of manipulating a alone and allowing the mediator to take its natural value under a. In both cases, we make the standard Stable Unit Treatment Value assumption (SUTVA) such that there is only one version of treatment a and that each unit's potential outcomes depend only on their individual treatment assignments.[1] This also allows us to connect the observed outcomes to potential outcomes such that $M_i = M_i(a)$ and $Y_i = Y_i(a) = Y_i(a, M_i(a))$ for units observed to have $A_i = a$. Likewise $Y_i = Y_i(a, m)$ for units with observed $A_i = a$ and $M_i = m$.

[1] Angrist, Imbens, and Rubin 1996.

1.2 Identification of total effects

We start by briefly reviewing identification of the total effect of the treatment. For an individual unit, the effect of treatment A on outcome Y is defined as

$$\tau_i = Y_i(1) - Y_i(0)$$

Likewise, the individual effect of A on M is defined as

$$\mu_i = M_i(1) - M_i(0)$$

Since individual treatment effects are unidentifiable from the data as we only observe units under one treatment condition,[2] researchers typically focus on identifying and estimating averages of effects. Define the average treatment effects for Y and M as $\tau = E[\tau_i] = E[Y_i(1) - Y_i(0)]$ and $\mu = E[\mu_i] = E[M_i(1) - M_i(0)]$ respectively. Identifying either average effect from observational data requires two primary assumptions in addition to SUTVA: an ignorability or exogeneity assumption – often conditional on pre-treatment covariates X_i – and an "overlap" or positivity assumption.[3]

Assumption 1 *Ignorability of A_i*

$$\{Y_i(a, m), M_i(a)\} \perp\!\!\!\perp A_i | X_i = x$$

for all values of a and m

Assumption 2 *Positivity*

$$0 < Pr(A_i = a | X_i = x) < 1$$

for all values of a and x within the support of A_i and X_i.

Under these assumptions, the difference in conditional expectations identifies the conditional treatment effect and the average can be recovered by averaging over the empirical distribution of

[2] Holland 1986.

[3] Imbens 2004.

covariates.

$$E[Y_i(1) - Y_i(0)|X_i = x] = E[Y_i|A_i = 1, X_i = x] - E[Y_i|A_i = 0, X_i = x]$$
$$E[Y_i(1) - Y_i(0)] = \{E[Y_i|A_i = 1, X_i = x] - E[Y_i|A_i = 0, X_i = x]\}\, Pr(X_i = x)$$

When X_i is low-dimensional, each conditional expectation can be estimated via its sample analogue, yielding the sub-classification estimator. Alternatively, regression or matching estimators – among others – can be used to impute the unobserved potential outcomes.[4] If researchers are willing to assume a regression model for the outcome, the average treatment effect is identified by a regression coefficient under a constant effects assumption. Assume:

$$Y_i = \alpha_1 + \beta_1 A_i + \xi_1' X_i + \varepsilon_{i1}$$

where ε_{i1} is a mean-zero, finite variance error term. The β_1 coefficient denotes the average treatment effect of A.[5] However, note that when effects are not constant, multiple regression yields a variance-weighted average of individual treatment effects.[6]

1.3 Decomposing the total effect

In mediation analysis, researchers are typically interested in decomposing the average treatment effect τ into components mediated by and unmediated by the post-treatment quantity M. While a number of decompositions have been proposed in the statistics literature,[7] one of the most common is a two-part division of the average treatment effect τ into a "direct" effect (ξ) and an indirect effect (δ).[8] The "indirect" effect or Causal Mediation Effect (CME) Imai, Keele, Yamamoto, et al. 2010 for unit i corresponds to the difference in potential outcomes if unit i were assigned the value of its mediator under treatment versus the value if i were assigned the value of its mediator under control, fixing the value of the treatment to t.

Formally, the individual indirect effect is defined as

[4] Imbens 2004.

[5] This formulation of the outcome model is drawn from the linear structural model set-up in Imai et al. 2011.

[6] Aronow and Samii 2016.

[7] e.g. VanderWeele 2013, VanderWeele 2014

[8] Imai, Keele, Yamamoto, et al. 2010, Robins and Greenland 1992

$$\delta_i = Y_i(a, M_i(1)) - Y_i(a, M_i(0))$$

or the effect on Y of manipulating M_i to take on the value it would take under treatment compared to manipulating it to take on the value it would take under control.

Likewise, the individual direct effect is

$$\zeta_i = Y_i(1, M_i(a)) - Y_i(0, M_i(a))$$

for all values of m and x within the sup

or the effect of treatment fixing M_i to $M_i(a)$.

We can write the total effect as the sum of a direct effect and an indirect effect:

$$\tau_i = \delta_i(a) + \zeta_i(1-a)$$

and the average treatment effect as

$$\tau = E[\delta_i(a)] + E[\zeta_i(1-a)] = delta(a) + \zeta(1-a)$$

connection to the Baron-Kenny mediation analysis approach and

Identification and sensitivity analysis via sequential ignorability

Identification of the average direct and indirect effects requires additional assumptions beyond those required for identification of the average treatment effect. One approach is to assume sequential ignorability of the outcome and mediator.[9]

[9] Imai, Keele, Yamamoto, et al. 2010.

Assumption 3 *Sequential ignorability*

$$Y_i(a', m), M_i(a) \perp\!\!\!\perp A_i | X_i = x$$
$$Y_i(a', m) \perp\!\!\!\perp M_i(a) | A_i = a, X_i = x$$

where $0 < p(M_i(a) = m | A_i = a, X_i = x) < 1$, *for all values of* m *and* x *within the support of* M_i *and* X_i.

The first part of the assumption is the standard ignorability assumption for treatment. The second part states that the conditional on treatment and pre-treatment covariates, the mediator is as-good-as randomly assigned. This latter assumption is very strong and often difficult to validate as even in a randomized experiment it may not hold. It requires that the researcher account for all pre-treatment covariates that may be predictive of both the mediator and the outcome. Furthermore, it restricts researchers from conditioning on post-treatment quantities that may be consequences of treatment but confounders of the mediator and outcome.

To address the difficulty of justifying the second component of sequential ignorability, Imai, Keele and Yamamoto develop a sensitivity analysis technique that relaxes this assumption. The structural equation framework has been used to motivate analyses of causal mechanisms as far back as Baron and Kenny.[10] Imai, Keele and Yamamoto clarify the assumptions under which this framework can yield valid estimates of mediation effects.[11] They then motivate a sensitivity analysis that varies the correlation between the linear model error terms, though subsequent work has generalized this approach to other models.[12] We focus here on the linear structural equation modeling approach, both due to its connection to the older Baron-Kenny mediation analysis and its common use within political science.[13]

We retain the following model for the outcome and treatment alone.

$$Y_i = \alpha_1 + \beta_1 A_i + \xi_1' X_i + \varepsilon_{i1} \tag{1}$$

[10] Baron and Kenny 1986.

[11] Imai, Keele, Yamamoto, et al. 2010.

[12] Imai, Keele, and Tingley 2010.

[13] Imai et al. 2011.

We further assume the following model for the mediator

$$M_i = \alpha_2 + \beta_2 A_i + \xi_2' X_i + \varepsilon_{i2} \tag{2}$$

and another model for the outcome conditional on the mediator

$$Y_i = \alpha_3 + \beta_3 A_i + \gamma M_i + \xi_3' X_i + \varepsilon_{i3} \tag{3}$$

where ε_{i2} and ε_{i3} are mean zero error terms. Estimating both equations using least squares and taking the product $\hat{\beta}_2\hat{\gamma}$ yields the product-of-coefficients mediation effect.[14]. Imai et. al. show that this approach identifies the causal mediation effect under the assumption that the models are correctly specified along with an additional no-interaction assumption $\delta = \delta(0) = \delta(1)$.[15] They further note that this no-interaction assumption is not necessary if researchers specify a more flexible model for Y_i by including a treatment-mediator interaction. For simplicity, we focus on the case where the no-interaction assumption holds.

To relax the strong assumption of sequential ignorability with respect to the mediator, Imai et. al. propose a sensitivity parameter for the error correlation $\text{Cor}(\varepsilon_{i2}, \varepsilon_{i3}) = \rho$ where $-1 < \rho < 1$. When ρ is zero, sequential ignorability holds. Conversely, a non-zero ρ captures the presence of some unobserved omitted confounder of the mediator M and Y. The closer the correlation is to 1 or -1, the larger the magnitude of confounding. For any fixed value of ρ, and under a no-interaction assumption, the average causal mediation effect is identified by

$$\delta(1) = \delta(0) = \frac{\beta_2\sigma_1}{\sigma_2}\{\tilde{\rho} - \rho\sqrt{(1-\tilde{\rho}^2)/(1-\rho^2)}\}$$

where $\tilde{\rho} = Cor(\varepsilon_{i1}, \varepsilon_{i2})$ and $\sigma_j^2 = Var(\varepsilon_{ij})$. The variance terms can be estimated directly from the regressions regressions 1 and 2.

Notably, this does not address violations of the first part of sequential ignorability – the assumption that the treatment is randomly assigned given the covariates. We likewise leave this for other work, but note that it may be feasible to incorporate a number existing methods for

[14] MacKinnon et al. 2002.

[15] Imai, Keele, Yamamoto, et al. 2010; Imai et al. 2011.

conducting a sensitivity analysis of treatment exogeneity in a linear equation framework.[16]

Identification and sensitivity analysis via instrumental variables

Even with a sensitivity analysis, sequential ignorability may be difficult to justify in a research context such as when post-treatment confounding exists between M and Y. Additionally, the sensitivity parameter ρ may be difficult for researchers to interpret. Imai et. al. suggest a re-parameterization in the vein of Imbens which allows researchers to specify the sensitivity parameter in terms of the share of unexplained variance explained by an unobserved confounder.[17] This gives a quantity that captures the overall strength of the confounder's effect on both the treatment and the mediator in a standardized fashion. However, even in this situation, reasonable benchmarks for how strong a likely unobserved confounder might be are difficult to formulate. Furthermore, this re-parameterization requires researchers to specify two sensitivity parameters – one relating to the confounder's effect on the mediator and the other relating to its effect on the outcome.

This paper presents an alternative strategy for both identification and sensitivity analysis. It builds on an argument found in Sobel 2008 that an instrumental variables approach identifies mediation effects. It notes that these assumptions are not necessarily weaker than those required for identification. While it does not require sequential ignorability, it requires additional restrictions on the paths through which the treatment affects the mediator and outcome consistent with the idea that the treatment acts as an instrument for the mediator. However, by suggesting a sensitivity analysis on some of the more implausible assumptions – namely, the "exclusion restriction" – we develop a method that may be more useful to applied researchers in some circumstances. In particular, we believe that researchers are more likely to be able to conceive of reasonable bounds for a sensitivity parameter that reflects the hypothesized "direct effect" of the treatment.

We will first discuss identification of the mediation effect via instrumental variables in the context of no covariates. We will then extend it to the linear structural model case discussed in the previous section and explain the proposed sensitivity analysis and how it differs from the sensitivity analysis for identification under sequential ignorability.

We will start with the binary mediator/binary treatment case with no covariates. In this

[16] Imbens 2003; Cinelli and Hazlett 2018; Oster 2019.

[17] Imai, Keele, Yamamoto, et al. 2010; Imbens 2003.

situation the average treatment effect τ and the treatment effect on the mediator are identified by a simple difference in means in outcomes across treatment arms. To obtain the effect of the mediator from these two effects, we introduce the instrumental variables assumptions from Angrist, Imbens and Rubin.[18] First, we assume that the treatment effect of A on M is monotonic for all units

Assumption 4 *Monotonicity*

$$M_i(1) > M_i(0) \ \forall \ i$$

This assumption partitions the sample into three sub-groups based on their joint potential outcomes $M_i(1), M_i(0)$. [19]. Units can have either $M_i(1) = M_i(0) = 1$, $M_i(1) = M_i(0) = 0$ or $M_i(1) = 1, M_i(0) = 0$. Let $S_{ab} = M_i(1) = a, M_i(0) = b$ denote the stratum characterized by the joint potential outcomes $M_i(1), M_i(0)$. For example, S_{10} denotes the sub-group with $M_i(1) = 1, M_i(0) = 0$.[20]

Second, instrumental variables designs make an "exclusion restriction" assumption which rules out the presence of a direct effect of A on the individual level.

Assumption 5 *Exclusion Restriction*

$$Y_i(a, m) = Y_i(a', m)$$

for all m, a, a'

The average treatment effect can be decomposed as follows:

$$\begin{aligned}\tau = E[Y_i(1) - Y_i(0)] = E[Y_i(1, M_i(1)) - Y_i(0, M_i(0))] &= E[Y_i(1, M_i(1)) - Y_i(0, M_i(0))|S_{11}]Pr(S_{11}) \\ &+ E[Y_i(1, M_i(1)) - Y_i(0, M_i(0))|S_{00}]Pr(S_{00}) \\ &+ E[Y_i(1, M_i(1)) - Y_i(0, M_i(0)))|S_{10}]Pr(S_{10})\end{aligned}$$

[18] Angrist, Imbens, and Rubin 1996.

[19] These are often referred to as "principal strata" with respect to the mediator. See Frangakis and Rubin 2002.

[20] Note that by monotonicity, S_{01} is ruled out.

Substituting $M_i(1)$ for the values implied by each stratum yields:

$$\begin{aligned}\tau &= E[Y_i(1,1) - Y_i(0,1)|S_{11}]Pr(S_{11})\\ &+ E[Y_i(1,0) - Y_i(0,0)|S_{00}]Pr(S_{00})\\ &+ E[Y_i(1,1) - Y_i(0,0)|S_{10}]Pr(S_{10})\end{aligned}$$

Note that the treatment effects for the two non-complier strata are direct effects only, while the "complier" stratum effect further decomposes into an indirect and direct effect.

$$\begin{aligned}\tau &= E[Y_i(1,1) - Y_i(0,1)|S_{11}]Pr(S_{11})\\ &+ E[Y_i(1,0) - Y_i(0,0)|S_{00}]Pr(S_{00})\\ &+ E[Y_i(1,1) - Y_i(0,1)|S_{10}]Pr(S_{10})\\ &+ E[Y_i(0,1) - Y_i(0,0)|S_{10}]Pr(S_{10})\end{aligned}$$

Under the "exclusion restriction", the first three terms all equal 0, yielding

$$\tau = E[Y_i(0,1) - Y_i(0,0)|S_{10}]Pr(S_{10})$$

Since $Pr(S_{10}) = E[M_i(1) - M_i(0)]$ under monotonicity, the treatment effect reduces to the product of the average effect of A on M multiplied by the average effect of M on Y within the "complier" stratum.

$$\tau = E[Y_i(0,1) - Y_i(0,0)|S_{10}] \times E[M_i(1) - M_i(0)]$$

We can further show that this is equivalent to the pure indirect effect or average causal mediation effect (ACME) from Imai, Keele and Yamamoto $\delta(a)$.[21]

[21] Imai, Keele, Yamamoto, et al. 2010.

$$\begin{aligned}\delta(a) = E[Y_i(a, M_i(1)) - Y_i(a, M_i(0))] &= E[Y_i(1, M_i(1)) - Y_i(0, M_i(0))|S_{11}]Pr(S_{11}) \\ &+ E[Y_i(1, M_i(1)) - Y_i(0, M_i(0))|S_{00}]Pr(S_{00}) \\ &+ E[Y_i(1, M_i(1)) - Y_i(0, M_i(0)))|S_{10}]Pr(S_{10})\end{aligned}$$

Substituting in the values of $M_i(1), M_i(0)$ for each stratum yields

$$\begin{aligned}\delta(a) = E[Y_i(a, M_i(1)) - Y_i(a, M_i(0))] &= E[Y_i(a, 1) - Y_i(a, 1)|S_{11}]Pr(S_{11}) \\ &+ E[Y_i(a, 0) - Y_i(a, 0)|S_{00}]Pr(S_{00}) \\ &+ E[Y_i(a, 1) - Y_i(a, 0)|S_{10}]Pr(S_{10}) \\ &= E[Y_i(a, 1) - Y_i(a, 0)|S_{10}]Pr(S_{10})\end{aligned}$$

Notably, the exclusion restriction also implies a no-interaction assumption with respect to the indirect effect such that $\delta(0) = \delta(1)$. Therefore, under the instrumental variables assumptions, the reduced-form effect of A on Y is equivalent to the natural indirect effect. A similar point is made in Appendix B of Imai, Keele and Tingley.[22].

However, we may not be strictly interested in the mediation effect for the entire sample. Rather, we are interested in knowing whether there exists a mediation effect *among those units for which the treatment shifted the mediator*. We can then rescale the pure indirect effect to the average effect of the mediator by dividing the average treatment effect by the effect of treatment on the mediator, which is the standard Wald ratio estimator for instrumental variables.[23]

$$\frac{E[Y_i(1) - Y_i(0)]}{E[M_i(1) - M_i(0)]} = E[Y_i(0, 1) - Y_i(0, 0)|S_{10}]$$

As Angrist and Imbens illustrate, the effect identified is the average effect of the mediator on the outcome for the stratum of units for which treatment induces a change in the mediator (the compliers). Therefore, under the instrumental variables assumptions for the treatment, the

[22] Imai, Keele, and Tingley 2010.

[23] Angrist, Imbens, and Rubin 1996.

IV estimator identifies the average causal mediation effect among the stratum of units who are affected by the treatment to change the mediator.

Suppose then that we relaxed the exclusion restriction but assumed that the direct effects within each stratum are constant across mediator levels

Assumption 6 *No mediator-treatment interactions for direct effects*

$$E[Y_i(1,0) - Y_i(0,0)|S_{ab}] = E[Y_i(1,1) - Y_i(0,1)|S_{ab}]$$

for $a = \{0,1\}$, $b = \{0,1\}$

We can then write the average treatment effect in terms of the indirect effect of interest plus the sum of three direct effects across each of the three principal strata

$$\tau = \rho + E[Y_i(0,1) - Y_i(0,0)|S_{10}] \times E[M_i(1) - M_i(0)]$$

where

$$\begin{aligned}\rho &= E[Y_i(1,1) - Y_i(0,1)|S_{11}]Pr(S_{11}) \\ &+ E[Y_i(1,0) - Y_i(0,0)|S_{00}]Pr(S_{00}] \\ &+ E[Y_i(1,1) - Y(1,0)|S_{10}]Pr(S_{10})\end{aligned}$$

The no-interaction assumption allows us to further simplify :

$$\begin{aligned}\rho &= E[Y_i(1,1) - Y_i(0,1)|S_{11}]Pr(S_{11}) + E[Y_i(1,1) - Y_i(0,1)|S_{00}]Pr(S_{00}) + E[Y_i(1,1) - Y(1,0)|S_{10}]Pr(S_{10}) \\ &= E[Y_i(1,1) - Y_i(0,1)]\end{aligned}$$

where the second step follows from the monotonicity assumption implying $Pr(S_{11}) + Pr(S_{10}) + Pr(S_{00}) = 1$.

Therefore, under a no treatment-mediator interaction assumption, we can decompose the average treatment effect into the quantity of interest and a sensitivity parameter reflecting the average direct effect of treatment holding the mediator fixed. Even if we do not make the no-interaction assumption, the parameter ρ represents a stratum-weighted average of direct effects. However in this case, we are averaging over two types of effects ($Y_i(1,1) - Y_i(0,1)$ and $Y_i(1,0) - Y_i(0,0)$) in different proportions and the interpretation of the sensitivity parameter becomes more difficult. While it is possible to set multiple sensitivity analysis parameters for each of these effects and combine them appropriately, we consider a single parameter sufficient for most applications.

Given a fixed violation of the exclusion restriction, ρ, the effect of the mediator for the complier stratum can be identified by subtracting the sensitivity analysis parameter from the treatment effect and dividing by the effect of treatment on the mediator.

$$\frac{E[Y_i(1) - Y_i(0)] - \rho}{E[M_i(1) - M_i(0)]} = E[Y_i(0,1) - Y_i(0,0)|S_{10}]$$

We can then consider varying ρ across a range of feasible values to assess how the implied effect of the mediator on the outcome changes as the researchers assumptions regarding the direct effect change. Notably, we have made no assumptions regarding confounding of M and Y, allowing for possible unobserved variables to affect both.

To extend the analysis to cases where the mediator and treatment are not binary and identification is conditional on covariates, we adopt the same linear structural model set-up from Imai et. al. and use the standard two-stage least squares (2SLS) approach to estimate the effect of the mediator.[24] Since linear regression models are commonly used for covariate adjustment in most observational studies, these modeling assumptions are already implicit in a large share of studies of counteraction - including those we analyze in the main text. One important caveat is that structural models require strong assumptions regarding effect homogeneity to justify various product-of-coefficients techniques to estimate mediated effects irrespective of the other assumptions used to identify the effect of the mediator.[25]

We assume the following models for the outcome and mediator.

[24] Imai et al. 2011.

[25] See Glynn 2012 for examples of where these approaches can fail. The central intuition is that while products of coefficients correspond to indirect effects on an individual level, researchers can only estimate average effects. When treatment has an effect on the mediator for one subset of units which are radically different from those where the mediator affects on the outcome, these products can be misleading.

$$M_i = \alpha_2 + \beta_2 A_i + \xi_2' X_i + \varepsilon_{i2} \tag{4}$$

and another model for the outcome conditional on the mediator

$$Y_i = \alpha_3 + \beta_3 A_i + \gamma M_i + \xi_3' X_i + \varepsilon_{i3} \tag{5}$$

where ε_{i2} and ε_{i3} are error terms that are allowed to be correlated. The conditional ignorability of A_i implies that $E[\varepsilon_{i2}|A_i, X_i] = 0$. However, because M_i is not exogenous conditional on A and X, $E[\varepsilon_{i3}|M_i, A_i, X_i] \neq 0$. However, we do have $E[\varepsilon_{i3}|A_i, X_i] = 0$ by ignorability of A_i. β_2 denotes the average treatment effect of a unit increase in a on M_i. Likewise, *gamma* denotes the average treatment effect of a unit increase in m on Y_i holding constant A_i. Assuming a constant β_2 implicitly makes an assumption stronger than monotonicity as the effect is not simply positive or negative for all units, but a constant α.

We are interested in estimating the parameter γ, denoting the effect of the mediator on the outcome. However, estimating the second regression via OLS alone will yield biased estimates of γ. Under the exclusion restriction, $\beta_3 = 0$. Therefore, we can write

$$Y_i = \alpha_3 + \gamma M_i + \xi_3' X_i + \varepsilon_{i3} \tag{6}$$

The regression of M on A and X can be estimated via the standard least squares approach, yielding coefficient estimates $\hat{\alpha_2}$, $\hat{\beta_2}$, $\hat{\xi_2'}$. Substituting those fitted values in the second regression yields

$$Y_i = \alpha_3 + \gamma \left[\hat{\alpha_2} + \hat{\beta_2} A_i + \hat{\xi_2'} X_i\right] + \xi_3' X_i + \varepsilon_{i3} \tag{7}$$

$$Y_i = \alpha_3 + \gamma\hat{\alpha_2} + \gamma\hat{\beta_2} A_i + \left[\hat{\xi_2'} + \xi_3'\right] X_i + \varepsilon_{i3} \tag{8}$$

Since $E[\varepsilon_{i3}|A_i, X_i] = 0$, regressing Y_i on the fitted values from the first regression and the covariates yields identifies the parameter γ.

We motivate the sensitivity analysis technique by assuming that the violation of the exclusion restriction is equal to ρ. This allows us to state that the exclusion restriction will hold for a transformed outcome: $Y_i - \rho A_i$. This then allows us to write

$$Y_i - \rho A_i = \alpha_3 + \gamma M_i + \xi_3' X_i + \varepsilon_{i3} \tag{9}$$

and use the two-stage least squares estimator to estimate γ.

While this sensitivity analysis is motivated using the same linear structural equation models as in the work by Imai et. al., it differs in two key respects. The Imai et. al. sensitivity analysis framework starts from the assumption that researchers have accounted for all confounders of M and Y in X and the sensitivity parameter relaxes this assumption by capturing the magnitude of omitted confounding between M and Y. Under the approach suggested here, there are no conditional independence assumptions made with respect to M and Y. Rather, the instrumental variables assumptions hinge on the absence of a direct effect from A to Y that does not flow through M, which is relaxed by varying the magnitude of the exclusion restriction. The Imai et. al. sensitivity analysis varies the correlation between error terms in the model, the sensitivity analysis we suggest varies the direct effect of treatment. Furthermore, we assume that researchers are interested in both the average causal mediation effect (the combination of the $A \rightarrow M$, and $M \rightarrow Y$ pathways) and the average effect of M on Y. While it is possible to conduct a sensitivity analysis just for the first by regressing outcome on the treatment, giving an estimate of $\hat{\beta}_2\gamma$, this does not permit researchers to distinguish between cases where treatment affects the mediator for very few units but γ is larger and cases where treatment affects many units but γ is small. In assessing counteraction hypotheses in particular, we recommend researchers evaluate both the effect of A on M and the effect of M on Y separately. In the instrumental variables design, the latter is identified particularly for the sub-group of interest: those units for which A moves M.

2 Replication details

2.1 Hafner-Burton, 2008

We make a few modifications to the Hafner-Burton data in our replication. Unfortunately, we are unable to replicate the results from the original paper exactly as presented, even when utilizing data that appears to be as close as possible to the paper's original construction. Moreover, we noticed substantial missingness in the original dataset with respect to key confounding variables - in particular, population. We corrected these issues and updated the data to cover all countries coded in the CIRI human rights data from inception (1981) to 2000, when the paper's covariates conclude. Since we have CIRI measures for 2001 and the original analyses used lagged outcomes, we can also include 2001 in our analysis. Overall, we are able to report a total of 2173 complete observations in the period from 1984 to 2001 (allowing for a three-period lag for the CIRI rights outcome as suggested in the original paper).

Among the other corrections we make is to analyze comparable time periods for the CIRI data and the political rights index reported in the original paper. As Table 2 of Hafner-Burton shows, the reported sample size for the regressions on the political rights outcome is notably larger than the sample size for the political terror regressions. We find that this is driven by differences in coverage rates (as CIRI only begins in 1981). As we note in the main text, it is important to analyze treatment effects for comparable samples in order to make claims about mediation and counteraction. We also re-code the NGO shaming variable to have a more intuitive interpretation. The original paper uses the (de-meaned) log of the number of mentions of a given country by Amnesty International in a particular year. Because for many years, this value is equal to 0 and the log of 0 is undefined, the paper replaces these zeroes with a small value. While this is not stated anywhere explicitly in the original paper, it is implied that this value is .1 based on the minimum value given in the summary statistics for the advocacy variable in Table 1 ($\log(.1) = -2.303$). Since the choice of this value is arbitrary and results in an uninterpretable scale for the magnitude of the treatment effects, we consider an alternative coding that coarsens the presence or absence of shaming into a binary indicator - whether Amnesty International issued any press releases regarding that country in a given year. This appears to be the case in roughly one-third of all country-years in our dataset. This indicator correlates at about .5 with the logged measure in the original paper and behaves similarly. Analyses with the original logged measure are qualitatively identical to those we present here and still fail to recover a statistically significant effect for Amnesty International shaming under the author's originally specified

	Terror	Terror	Terror	Political
(Intercept)	0.01	0.00	−0.02	0.81^{***}
	(0.31)	(0.34)	(0.42)	(0.23)
Shaming (lag 1 year)	0.11	0.12^{*}	0.18^{**}	0.03
	(0.07)	(0.07)	(0.08)	(0.03)
Repression (lag 1 year)	0.43^{***}	0.48^{***}	0.66^{***}	0.89^{***}
	(0.03)	(0.02)	(0.02)	(0.02)
Repression (lag 2 years)	0.17^{***}	0.27^{***}		
	(0.03)	(0.02)		
Repression (lag 3 years)	0.18^{***}			
	(0.02)			
CAT (lag 1 year)	−0.02	−0.01	−0.00	−0.04
	(0.06)	(0.06)	(0.07)	(0.03)
CCPR (lag 1 year)	0.06	0.05	0.05	-0.06^{*}
	(0.07)	(0.07)	(0.08)	(0.04)
Democracy (lag 1 year)	-0.27^{***}	-0.28^{***}	-0.35^{***}	-0.21^{***}
	(0.08)	(0.08)	(0.08)	(0.07)
Log GDP per capita (lag 1 yr)	-0.11^{***}	-0.13^{***}	-0.17^{***}	-0.04^{***}
	(0.02)	(0.02)	(0.03)	(0.01)
Log Population (lag 1 year)	0.11^{***}	0.13^{***}	0.16^{***}	−0.00
	(0.02)	(0.02)	(0.02)	(0.01)
Civil war (lag 1 year)	0.30^{**}	0.35^{***}	0.53^{***}	0.06
	(0.12)	(0.12)	(0.13)	(0.04)
War (lag 1 year)	0.39^{***}	0.43^{***}	0.43^{**}	0.11^{*}
	(0.13)	(0.15)	(0.18)	(0.06)
Fixed effects for year	Yes	Yes	Yes	Yes
N	2173	2173	2173	2173
N Countries	140	140	140	140

OLS coefficients. Cluster-robust (country) standard errors in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 1: Re-analysis of Hafner-Burton (2008): Effect of Amnesty International shaming on physical and political rights

Table 1 presents coefficient estimates for the regression models predicting physical integrity rights and political rights.[26] The original paper argues for a three-period lag model for physical integrity rights and a one-period lag for political rights. When re-estimating these models on our

[26] These reflect the estimates for Amnesty International shaming from Table 2 in the original paper.

updated dataset, the direction of the physical integrity rights estimates are consistent with those presented in the original paper. By relaxing the three-period lag assumption, we obtain stronger estimated effects for shaming, but this is likely being driven by omitted variable bias as we find the third and second lags remains a strong predictor of the outcome. Regardless of the choice of covariates for the terror indices, the absence of any meaningful effect of shaming on political rights on a comparable sub-sample is evidence against counteraction.

2.2 Lupu, 2013

In the case of Lupu, our reconstructed dataset contains slightly more observations (2155 versus 1966, similar to our replication of Hafner-Burton). Additionally, the original paper estimated ordinal probit regression models for each of the civil and personal integrity rights indicators. Instead of fitting ordinal probit models, we fit a standard ordinary least-squares regression of the CIRI scores. While the scores are not properly "continuous," OLS still returns consistent estimates of the conditional expectation function. The coefficients from the OLS models also have direct interpretations as changes in the conditional expectation of the mediator and outcome. Additionally, linear models are required for using the fitted values approach in two-stage least squares. Regardless of the change in the model, the estimates are all roughly comparable and coefficients are in the same direction.

Table 2 replicates the relevant ordinal probit models from Lupu's Table 3 and 4 for ICCPR ratification. The coefficient estimates are very close to those reported in the original paper and the significance thresholds are roughly the same for each variable, though notably, we do not recover statistically significant effects for speech or religious freedom outcomes. Table 3 presents the corresponding OLS estimates that we use in the main paper. Again, the directionality of the coefficients largely matches what was presented in the original paper.

	Disappearances	Association	Speech	Religious Freedom
ICCPR Ratification	-0.16^{*}	0.22^{**}	0.11	0.10
	(0.08)	(0.09)	(0.09)	(0.08)
Judicial independence	0.18^{**}	0.13^{**}	0.21^{***}	0.19^{***}
	(0.07)	(0.06)	(0.07)	(0.05)
Polity	-0.00	0.08^{***}	0.09^{***}	0.04^{***}
	(0.01)	(0.01)	(0.01)	(0.01)
Regime durability	0.01^{*}	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)
Civil war	-0.85^{***}	-0.15	-0.23^{***}	-0.06
	(0.09)	(0.10)	(0.09)	(0.11)
External war	0.04	-0.05	-0.14	-0.26
	(0.21)	(0.20)	(0.15)	(0.17)
GDP per capita (logged)	-0.00	-0.06^{*}	0.04	-0.12^{***}
	(0.03)	(0.03)	(0.03)	(0.03)
Population (logged)	-0.09^{***}	-0.06^{***}	-0.03^{**}	-0.13^{***}
	(0.01)	(0.01)	(0.01)	(0.01)
INGOs	0.00	0.00	-0.00	0.00^{*}
	(0.00)	(0.00)	(0.00)	(0.00)
Rights (t-1)	1.07^{***}	1.61^{***}	1.03^{***}	1.21^{***}
	(0.09)	(0.10)	(0.08)	(0.06)
Fixed effects for year	Yes	Yes	Yes	Yes
N	2155	2155	2155	2155
N Countries	144	144	144	144

Ordered probit coefficents. Cluster-robust (country) standard errors in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 2: Replication of Lupu (2013) - Ordinal Probit Regressions

	Disappearances	Association	Speech	Religious Freedom
(Intercept)	1.02^{***}	0.63^{***}	0.54^{**}	1.48^{***}
	(0.22)	(0.18)	(0.27)	(0.30)
ICCPR Ratification	-0.05^{**}	0.06^{**}	0.04	0.04
	(0.02)	(0.03)	(0.03)	(0.04)
Judicial independence	0.06^{**}	0.04^{**}	0.08^{***}	0.08^{***}
	(0.02)	(0.02)	(0.02)	(0.02)
Polity	-0.00	0.03^{***}	0.03^{***}	0.02^{***}
	(0.00)	(0.00)	(0.00)	(0.00)
Regime durability	0.00	-0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)
Civil war	-0.41^{***}	-0.05	-0.08^{***}	-0.02
	(0.06)	(0.03)	(0.03)	(0.04)
External war	0.01	-0.00	-0.05	-0.11^{*}
	(0.08)	(0.06)	(0.05)	(0.06)
GDP per capita (logged)	-0.00	-0.02	0.01	-0.05^{***}
	(0.01)	(0.01)	(0.01)	(0.02)
Population (logged)	-0.02^{*}	-0.01	-0.01	-0.05^{***}
	(0.01)	(0.01)	(0.01)	(0.01)
INGOs	0.00	0.00	-0.00	0.00^{*}
	(0.00)	(0.00)	(0.00)	(0.00)
Rights (t-1)	0.54^{***}	0.64^{***}	0.39^{***}	0.62^{***}
	(0.05)	(0.03)	(0.03)	(0.03)
Fixed effects for year	Yes	Yes	Yes	Yes
N	2155	2155	2155	2155
N Countries	144	144	144	144

OLS coefficients. Cluster-robust (country) standard errors in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 3: Replication of Lupu (2013) - Ordinary least squares regressions